Representation errors and retrievals in linear and nonlinear data assimilation

RMetS
Royal Meteorological Society

Representation errors and retrievals in linear and nonlinear data assimilation

Peter Jan van Leeuwen
Data Assimilation Research Centre, Department of Meteorology, University of Reading, UK

*Correspondence to: P. J. van Leeuwen, Department of Meteorology, University of Reading, Earley Gate, Reading RG6 6BB, UK.
E-mail: p.j.vanleeuwen@reading.ac.uk

This article shows how one can formulate the representation problem starting from Bayes' theorem. The purpose of this article is to raise awareness of the formal solutions, so that approximations can be placed in a proper context. The representation errors appear in the likelihood, and the different possibilities for the representation of reality in model and observations are discussed, including nonlinear representation probability density functions. Specifically, the assumptions needed in the usual procedure to add a representation error covariance to the error covariance of the observations are discussed, and it is shown that, when several sub-grid observations are present, their mean still has a representation error; so-called 'superobbing' does not resolve the issue. Connection is made to the off-line or on-line retrieval problem, providing a new simple proof of the equivalence of assimilating linear retrievals and original observations. Furthermore, it is shown how nonlinear retrievals can be assimilated without loss of information. Finally we discuss how errors in the observation operator model can be treated consistently in the Bayesian framework, connecting to previous work in this area.

Key Words: data assimilation; representation errors; Bayes' theorem

Received 6 June 2014; Revised 27 August 2014; Accepted 16 September 2014; Published online in Wiley Online Library

1. Introduction

Representation errors have since long troubled scientists involved in data assimilation. One of the reasons is that no clear definition is used by the community. Several scientists (e.g. Thacker, 2003; Janjic and Cohn, 2006; Köhl *et al.*, 2007; Leeuwenburgh, 2007; Oke and Sakov, 2008) define the representation error as the component of the observation error due to unresolved scales. Others focus on the mis-specification of the observation operator (Lorenc, 1986; Liu and Rabier, 2002). More generally, representation errors have been defined as errors due to any physical processes appearing in the observations but not in the model (Anderson *et al.*, 2005; Zaron and Egbert, 2006; Ponte *et al.*, 2007). Note that 'model' in this context and article will mean the dynamical model. This meaning will be used throughout the article. Sometimes the errors in the model equations, either from missing physics or from discretisation, are taken as part of the representation error. (Also the nomenclature is not identical among different authors. Some call these errors representivity errors, others representativity; we will use the simpler representation errors.) Obviously, representation errors will depend on model resolution, as do model errors, but they enter the data assimilation problem differently.

In this article the distinction between model error and representation error is taken very strictly. Representation errors are defined here as resulting from different representations of reality in model and observations. Although the model will be an imperfect representation of reality, the data assimilation problem does not address that since the solution to the data assimilation problem is the posterior probability density function (pdf) of this model given the observations, not of reality given the observations. As such, only representation differences between model and observations come into play. With this definition, they enter the data-assimilation problem via the likelihood and not via the prior probability density of the model. Strictly speaking, the terminology 'error' is misleading, in that no error has been made, and it is also not related to 'small-scale noise'. It is rather a misfit term: model and observations describe different parts of reality. With this definition, it is stressed that we explicitly consider all errors in the model as part of the prior. The fact that the model cannot be trusted at its smallest scales does not change that.

There is another related issue which needs a brief discussion. If there is a representation error, one could argue that observations try to bring information into the model which push it out of its 'natural' state, e.g. push it into unbalanced states. In operational numerical weather forecasting, filtering operations are typically needed after each update step to filter out unwanted gravity waves which tend to ruin the forecast. The reasons are twofold. Firstly, if the atmosphere were linear, the gravity waves would do no or very little harm. In linear data assimilation methods (or,

more accurately, data assimilation methods in which the prior is assumed to be Gaussian), the prior covariance defines the linear space in which the updates can take place. If one would allow for a prior covariance evolving with the flow, like in a full Kalman filter, the projection of the update on the gravity waves would be small in a linear system, and not damaging. However, we cannot afford to evolve full covariances, so either we use approximations, such as in the Ensemble Kalman filter, or we use prescribed covariances, as in four-dimensional variational assimilation (4D-Var), or hybrids of the two. If the actual model is linear, the balances are known and prescribed projection onto the slow manifold is relatively straightforward. A low-rank approximation of the covariance will need an extra filtering step.

Secondly, the issue is to a large extent related to the fact that linear data assimilation methods are used on nonlinear systems. The forecasts are damaged due to the nonlinear coupling of gravity waves with (e.g.) precipitation. For a nonlinear model, the nonlinear balances are not known, or very hard to implement, and projection onto gravity waves is much more of a problem. Several filtering techniques are used in the covariance model to reduce the damage. Even fully nonlinear data assimilation methods could become unbalanced by the update step. An exception is a standard particle filter with resampling in which only the relative weights are adjusted by the observations, and the members are unchanged. However, the number of ensemble members needed would be astronomical (e.g. Snyder *et al.*, 2008; van Leeuwen, 2009). More efficient particle filters could have balance issues simply because the observations do not constrain all possible solutions (e.g. van Leeuwen, 2010; van Leeuwen, 2011). At this moment it is unclear how large this effect will be, and how to solve it if it is large. In this article we do not consider this kind of 'error' to be part of the representation error as defined above.

Typically representation errors are taken into account by inflating the error covariance of the observations, sometimes dependent on position and/or flow regime, or by adding an extra error covariance to the observation-error covariance in the likelihood (e.g. Derber and Rosati, 1989; Oke *et al.*, 2005; Rogel *et al.*, 2005; Schiller *et al.*, 2008). This is typically based on Lorenc (1986) also Daley (1993) who showed that, for Gaussian distributed variables, errors related to an inexact observation operator H can be added directly to observation-error covariance related to measurement noise. It is important to realise that Lorenc actually treats a different problem as will be discussed later (section 7). We have to understand the justification for this procedure, and how general it is. To this end, we will recapitulate the actual meaning of the likelihood, after which we discuss different methods to take representation errors into account.

This article is organised as follows. The origin of the representation error is discussed in the next section, followed by the different forms of representation mismatch and how to formally solve them in observation (section 4) and model (section 5) space. We touch upon the retrieval problem here as a good example of some of the issues in section 6. Section 7 discusses the situation when H is known approximately, and section 8 summarises and concludes the article.

2. Bayes' theorem and the representation error

At the heart of data assimilation is Bayes' theorem, which states that the posterior pdf can be obtained by multiplying the prior pdf with the likelihood:

$$p(x|y) = \frac{p(y|x)\,p(x)}{p(y)}. \tag{1}$$

The meaning of the terms is as follows:

(1) $p(x|y)$ is the posterior pdf, which is the probability density value of model state x given a set of observations y, for all possible model states. This pdf is the solution to the data assimilation problem, and one could argue that Bayes' theorem shows that data assimilation is in fact a multiplication problem and not an inverse problem.
(2) $p(y|x)$ is the probability density value of the observation set y given that the model state is x, the so-called likelihood, i.e. the likelihood of the observations being y given that the state is x, for each of the possible model states.
(3) $p(x)$ is the prior probability density value of the model state x, for each possible model state, before observations have been taken into account.
(4) $p(y)$ is the prior probability density of the observations, i.e. before we know any of the possible states, and before we know the truth. It is not the probability density value of the observation given the truth, since the truth is also a model state (plus extra information on e.g. smaller scales). In fact 'nothing is given'. Since we use Bayes to evaluate the pdf of the state for a specific set of observations, $p(y)$ does not depend on x and is a normalisation constant that can quite often be ignored.

It is important to realise that Bayes' theorem is a point-wise equality, i.e. for each x, y combination the equality holds. In any specific data assimilation problem, the observation vector y is given, and the posterior pdf is only a function of the model state x. The above is completely general, and 'model state' can be changed to 'model parameters', or a combination of both. In all cases, the posterior pdf is obtained by multiplying the prior, our starting point, with the likelihood.

Before we continue, it is good to try to understand the likelihood term in case the model and the observations represent the true system in the same way, so when there is no representation error according to our definition. The observations are related to the true state of the system as

$$y = H_{\mathrm{t}}(x_{\mathrm{t}}) + \epsilon_{\mathrm{m}}, \tag{2}$$

in which H_{t} is the true observation operator, x_{t} is the true state of the system, and ϵ_{m} is a specific realisation of the measurement error. One could argue that, if the model is biased, there is still some sort of representation error that should be accounted for in the likelihood, as this bias will result in an extra model–observation misfit. However, that is incorrect. What we need in Bayes' theorem is not the likelihood of the observation given the true state of the system, but given any possible state x of the *modelled* system. So, the idea is that we measure that modelled state x, and then, using the distribution of the measurement error, determine the likelihood of the actual observation with value y. The fact that x is biased or obtained through a model with a systematic error is not relevant; this is the given state and we just need to know the likelihood of y given that state. Any bias term should appear in the prior. After this clarification, we now concentrate on the representation errors in the rest of this article.

3. The likelihood with exact observation operator

The representation error issue is related to the likelihood; that is where the observations come in, or rather, that is where model and observations are compared. The likelihood $p(y|x)$ is the value of the pdf of the observations y given that the model state is x. Several possible relations exist between observations and model states, and each leads to a different way to calculate this likelihood. In general the relation between model space and observation space is given by an operator H, so

$$y_{\mathrm{m}} = H(x) \tag{3}$$

transforms a model state x to vector y_{m} in observation space. y_{m} comes under several different names, modelled observation, forecasted observation, model equivalent of observation, etc. To increase confusion, the seismology community tends to call the H

operator the 'theory', and the parameter space (i.e. the slowness field) the 'model'.

Sections 3–6 focus on the case that H is known exactly. Section 7 deals with the case in which H is only known approximately.

Let us now have a look at the different possibilities for this H operator.

(1) The observations and the state represent the same phenomena. A special case is that in which there is a one-to-one relation between a model state variable and the model equivalent of the observation. In the more general case, several model variables are involved in producing a model equivalent of the observation, e.g. H can be an integral over a path in model space. Note that H can be nonlinear. This would restrict the data assimilation methods that can be used, but is not the focus of this article. This is an 'easy case' i.e. no extra information is needed to calculate the likelihood.

(2) The model has a higher level representation of reality than the observations. We have to find the closest relation between the model and the observation. For instance, the observations are low-resolution satellite observations, to be assimilated in a high-resolution model. In that case H is an averaging operator, i.e. a spatial average. Also this case is an 'easy case'.

(3) More problematic is the case when the observations represent phenomena that are not resolved by the model. An example is a point measurement, e.g. a temperature measurement with a thermometer, while the model has a spatial resolution of 1 km, or 100 km. To solve this case one has to realise what the likelihood actually means. $p(Y = y|X = x)$ is the probability that the observation random variable Y is realised as y given model state random variable $X = x$. The notation we use is that upper-case letters denote the random variable, and lower-case letters the realisation. The first thing to do is to find the relation between the pure observation and the closest model representation. There are two ways to approach this problem: via observation and via model space, as discussed below.

(4) Finally, observations and model can represent different phenomena. An example is a satellite measurement of the radiance at a certain wavelength, while the model has a temperature field, but no radiation-transfer code. This is usually treated using retrievals which transform the original observations to so-called retrieved variables which do have direct model equivalents. The question then is how one should assimilate these retrievals, one of the problems being that the retrieval is a pdf, not a specific value.

In the next sections we will concentrate on cases (3) and (4) above.

4. The likelihood via observation space

We discuss here the situation where the observation is of higher resolution but of the same type as (part of) the state vector.

To ease the presentation, we will assume that $H(x)$ is defined as the average over a certain modelled area i. The observation equation is given by:

$$y = H(x) + \epsilon \tag{4}$$

in which ϵ contains the representation and measurement errors, and $H(x)$ singles out that part of the model that relates best to the observation y.

This could be a model grid box, but, since the model has typically large errors at its finest resolution (think of e.g. group speed errors of waves), it could also be an average over several model grid boxes for several model variables. One should choose that combination of model variables that closely resembles the observation. So, from now on, $H(x)$ denotes this combination of model variables.

Assume the observation is located in this area i, then we need to calculate the probability of $Y = y$, given that the average value of y over this area is given. So we need

$$p(Y = y|\overline{Y} = \overline{y}) \tag{5}$$

and the likelihood becomes

$$p(y|\overline{y} = H(x)). \tag{6}$$

Intuitively it is clear that the width of this pdf must be larger then when the model can represent the observations. For the sake of argument, let us assume that the measurement errors are Gaussian distributed when the full state is given, with error covariance given by C_{m}. If only part of the full state is given as the model state, extra uncertainty is present in y which is not directly related to errors in the measurement process. This extra uncertainty is what is called the representation error. It should be realised, however, that there is no *a priori* reason to assume that the extra uncertainty leads to a Gaussian pdf, so assume an extra covariance can be added to C_{m}. This is illustrated below.

The question we want to answer is how to determine the pdf of Eq. (6). It is most straightforward to gather observations at more points in area i, and average them. When the average is equal (or close) to $H(x)$, we have one sample of $p(y|\overline{y} = H(x))$. We have to do a large number of these measurements to gather enough samples of $p\{y|\overline{y} = H(x)\}$ to be able to use it in Bayes' theorem, especially when we realise that this has to be done for each value of $H(x)$ that comes up, e.g. for each ensemble member in an Ensemble Kalman filter. Figure 1 illustrates this for one specific value of $H(x)$. Another way might be to make an 'educated guess', exploiting other prior knowledge on this pdf. This is what has to be done in nonlinear data assimilation.

We can formalise the above procedure as follows. Firstly, $H(x)$ is that part of the model state vector that is related to observations in area i. Next we define vector z in observation space which is related to the model state vector x as $z = H(x) + \widetilde{z}$. $\widetilde{z}$ is the high-resolution variation in area i, at the position of the observations. So, in our case z is a vector with elements $H(x) + \widetilde{z}$, in which $H(x)$ is the same for each element and $\widetilde{z}$ varies from element to element. These elements coincide with the observation locations.

From standard probability theory we obtain:

$$p(y|x) = \int p(y|x,\widetilde{z})\, p(\widetilde{z}|x)\, \mathrm{d}\widetilde{z} = \int p(y|z)\, p(\widetilde{z}|x)\, \mathrm{d}\widetilde{z}. \tag{7}$$

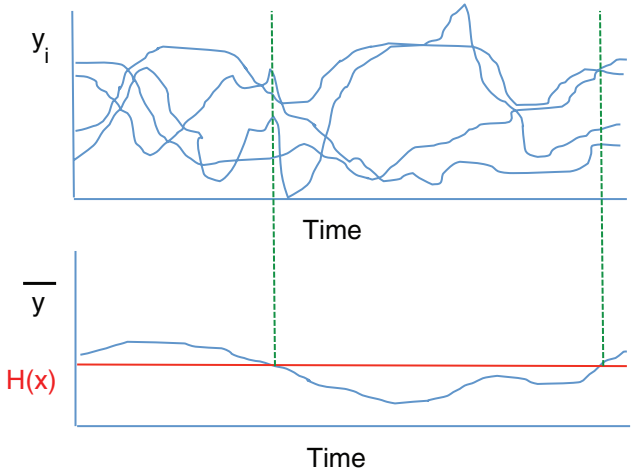

Figure 1. (a) shows individual observations and (b) the mean of these and $H(x)$. To calculate $p\{y|\overline{y} = H(x)\}$ or C_{r}, we take only those times into account when $\overline{y} = H(x)$, i.e. those times indicated by the dashed lines (green in online).

Now $p(y|z)$ is the pdf of the measurement errors, and these are assumed to be known, e.g. from the manufacturer of the measurement apparatus.

The second pdf $p(\widetilde{z}|x)$ is the pdf of high-resolution variations $\widetilde{z}$, given that the model average over area i is $H(x)$, and this is the factor responsible for the representation errors. We could use the measurement campaign and the method outlined above to find $p(\widetilde{z}|x)$.

To illustrate this further, assume both measurement errors and representation errors are Gaussian with error covariances C_m and C_r, respectively. When $H(x)$ is known and the model average is unbiased, the average over area i of $\widetilde{z} = 0$, so the Gaussian assumption leads to:

$$p(y|x) \propto \int \exp\left\{-\frac{1}{2}(y-z)^\mathrm{T} C_\mathrm{m}^{-1}(y-z) - \frac{1}{2}\widetilde{z}^\mathrm{T} C_\mathrm{r}^{-1}\widetilde{z}\right\} \mathrm{d}\widetilde{z}. \quad (8)$$

Assuming H is a linear operator and using $z = Hx + \widetilde{z}$, the integration over $\widetilde{z}$ can be performed, leading to

$$p(y|x) \propto \exp\left\{-\frac{1}{2}(y-Hx)^\mathrm{T}(C_\mathrm{m}+C_\mathrm{r})^{-1}(y-Hx)\right\}. \quad (9)$$

So, in this case the representation error appears as a covariance matrix C_r which has to be added to the measurement error covariance C_m to find the total observational error $R = C_\mathrm{m} + C_\mathrm{r}$. This derivation is the rationale behind the common way of adding the representation error covariance to the measurement error covariance. Note that, if the model average over area i is biased, we have that the average over $\widetilde{z} = z_0 \neq 0$. This leads, through a change of variable in the integral to $\widetilde{z} - z_0$, to a bias correction of the form $Hx \rightarrow Hx - z_0$ in the expression above.

The covariance C_r can be obtained from the set of measurements in area i. However, it is important to realise that the common procedure to use for C_r the covariance of the unconditioned observations y_i is not necessarily correct. The equations above show that one needs the covariance of the observations y_i given the mean state $H(x)$ over area i. The latter covariance will typically be smaller than the former for Gaussian distributed errors. This is illustrated in Figure 1. Also, since x varies over time, so will this covariance be time-dependent in general.

It is possible that the representation errors are correlated over time. The representation errors are due to the covariance of the unresolved scales conditioned on the mean state $H(x)$. Here we assume that the unresolved scales will decorrelate between observation times, also because the conditioning will change because the mean state $H(x)$ changes over time.

Obviously, when more observations are present in area i, all these m observations can be assimilated. If their errors are correlated this leads to the likelihood

$$p(y_1, ..., y_m|\overline{y} = H(x)), \quad (10)$$

in which $\overline{y}$ is the average over area i. Note that this has to be done for each possible model state x.

Another solution when we have the set of observations in area i is to average them and compare them directly to the model state x. This is sometimes called 'superobbing'. Will this elevate the representation error problem? The answer is no, as illustrated below. In this case likelihood is taken as:

$$p(\overline{y}|x). \quad (11)$$

To use this, we have to know the pdf of $\overline{y}$, and intuitively one might reason as follows. Assuming the y_i are independent with equal variance σ^2, in the limit of a large number of observations the central limit theorem allows us to approximate this pdf by a Gaussian distribution $N(\overline{y}, \sigma^2/m)$. However, it would be a mistake to assume that $\overline{y}$ has no representation error. This mean is just the mean of the observations, not the mean over the model area represented by $H(x)$. If we would like this mean to represent the mean of the model area, each individual observation has to be taken as an estimate of that model area. However, in that case one has to attach a representation error to each individual observation. So, as estimate of the model area, each observation is Gaussian distributed with error variance $\sigma_\mathrm{m}^2 + \sigma_\mathrm{r}^2$, leading to a combined estimate of the model area distributed as $N\{\overline{y}, (\sigma_\mathrm{m}^2 + \sigma_\mathrm{r}^2)/m\}$, under the assumption of independent identically distributed measurement and representation errors. *The main point is that one cannot get rid of the representation error.*

There is an interesting issue that the representation errors of the different observations y_i are likely to be correlated because all represent something different from the model in a similar way, related to e.g. missing model dynamics. Let us now for simplicity assume that the correlation in the representation errors between any two observations is given by ρ. In that case the variance of the observational mean including the representation error becomes

$$\sigma_{\overline{y}}^2 = \frac{\sigma_\mathrm{m}^2 + \sigma_\mathrm{r}^2}{m} + \frac{m-1}{m}\rho\sigma_\mathrm{r}^2. \quad (12)$$

The importance of this equation is that the correlated part of the representation error does not decrease with an increasing number of observations, so the representation error quickly becomes the dominant term in the error of the averaged observation. The main message is, again, that one cannot get rid of the representation error; it will always be present.

Let us now compare using the all observations individually in the data assimilation scheme to using the average observation. For the average observation, the update can be written as:

$$x_\mathrm{av} = x + (B^{-1} + H^\mathrm{T}\sigma_{\overline{y}}^{-2}H)^{-1}H^\mathrm{T}\sigma_{\overline{y}}^{-2}(\overline{y} - Hx), \quad (13)$$

where we used the information form of the Kalman filter update and again assumed H to be linear. The measurement operator Hx is the value of the state vector x in area i, which, just to remind the reader, should be a combination of model variables that closest represents the observation. For the set of observations we find an update:

$$x_\mathrm{d} = x + (B^{-1} + H_\mathrm{d}^\mathrm{T}R_\mathrm{d}^{-1}H_\mathrm{d})^{-1}H_\mathrm{d}^\mathrm{T}R_\mathrm{d}^{-1}(y - H_\mathrm{d}x), \quad (14)$$

in which $\mathrm{R_d}$ and $\mathrm{H_d}$ are the corresponding covariance and measurement operator for the set of observations. It is shown in the Appendix that this update is exactly equal to that of assimilating the average observation in Eq. (13).

However, this is only the case if we treat the average of the observations correctly. Ignoring the representation error in the average estimate is incorrect. It should be noted that the above equations are only true for Gaussian-distributed observation errors, and all observations have to be identically distributed with the same correlation between all observations. In the more general non-Gaussian case, the two solutions, assimilating individual observations versus assimilating their average, will lead to different results.

5. The likelihood via model space

More often than not, one does not have access to extra observations, so the problem is again how to compare the high-resolution observation to the low-resolution model. Several authors used high-resolution models to determine the relation between 'point' measurements and the larger-scale average in the model world, e.g. Janjic and Cohn (2006). It is important to understand what is done, and the analysis is similar to that given above, but the interpretation is slightly and importantly different. First we introduce a high-resolution model that has

the same resolution as the observations. If we denote the total high-resolution *model* variable by $z = (x, \widetilde{z})^{\mathrm{T}}$, where $\widetilde{z}$ denotes that part of the high-resolution model variable not represented by the coarse-resolution model variable x, we have

$$p(y|x) = \int p(y|x,\widetilde{z})\, p(\widetilde{z}|x)\, \mathrm{d}\widetilde{z} = \int p(y|z)\, p(\widetilde{z}|x)\, \mathrm{d}\widetilde{z}. \qquad (15)$$

For $p(y|z)$ we can use the pdf of the observations given the full state, and no representation error is present in this pdf, similar to the case discussed above. More difficult is $p(\widetilde{z}|x)$, the probability density of the high-resolution variations in the high-resolution model given a low-resolution average x.

One could build this pdf from a large number of model runs (or one very large model run assuming ergodicity), again similar to the observation-space solution, saving those states for which the low-resolution equivalent is (close to) x, Unfortunately, this has to be done for each x that comes up in the data assimilation problem, e.g. for an ensemble Kalman filter this has to be done for each ensemble member for each observation time.

Another way to calculate this pdf directly is to use a limited-area high-resolution model and use the same procedure as above, using only those states that have an area average close to x to generate the statistics for $p(\widetilde{z}|x)$. Since only a small proportion of the limited-area high-resolution model runs will have the desired area average, one could constrain the model runs to have the value x as mean over the model area in question. This limited-area high-resolution model can be considered part of the observation operator H. That procedure has the potential to strongly reduce the computational burden. Of course, the model has to be run long enough to gather enough information on $p(\widetilde{z}|x)$. However, imposing a model-area mean condition on the high-resolution model might not be straightforward since it will interfere directly with the model equations. A practical problem with both approaches is the boundary conditions on the high-resolution model. We might use an interpolated version of the low-resolution model, but it is unclear how this would affect the statistics of $p(\widetilde{z}|x)$. However, it is perhaps the best practical solution, the validity of which can be tested using a high-resolution model everywhere, and gathering statistics on $p(\widetilde{z}|x)$.

To make some progress analytically, let us first assume that all errors are Gaussian distributed, leading to

$$p(y|x) \propto \int \exp\left[-\frac{1}{2}\{y-H(z)\}^{\mathrm{T}} C_{\mathrm{m}}^{-1}\{y-H(z)\} - \frac{1}{2}\widetilde{z}^{\mathrm{T}} C^{-1}\widetilde{z}\right]\mathrm{d}\widetilde{z}, \qquad (16)$$

in which C_{m} denotes the covariance of the high-resolution variable $\widetilde{z}$ conditioned on the coarse-resolution variable $\mathbf{x}$. (Note that no bias correction is needed here by definition.) The following closely follows the arguments by Janic and Cohn (2006). It is noted that covariance C_{r} can depend on the value of x, so it can be time dependent, and is generally smaller than that of the unconstrained high-resolution model variability.

Let us now assume that H is linear, so that $Hz = \overline{H}x + \widetilde{H}\widetilde{z}$, which can easily be seen by realising that H is an (m, n_z) matrix, with m the number of observations, and n_z the size of the full state vector z, $\overline{H}$ is its (m, n_x) submatrix, and $\widetilde{H}$ is the rest of the H matrix, of size $(m, n_{\widetilde{z}})$. This leads to the likelihood

$$p(y|x) \propto \int \exp\left\{-\frac{1}{2}(y-\overline{H}x-\widetilde{H}\widetilde{z})^{\mathrm{T}} C_{\mathrm{m}}^{-1}(y-\overline{H}x-\widetilde{H}\widetilde{z}) - \frac{1}{2}\widetilde{z}^{\mathrm{T}} C^{-1}\widetilde{z}\right\}\mathrm{d}\widetilde{z}, \qquad (17)$$

which can be evaluated as

$$p(y|x) \propto \exp\left\{-\frac{1}{2}(y-\overline{H}x)^{\mathrm{T}}(C_{\mathrm{m}}+\widetilde{H}C\widetilde{H}^{\mathrm{T}})^{-1}(y-\overline{H}x)\right\}, \qquad (18)$$

showing how the representation error $C_{\mathrm{r}} = \widetilde{H}C\widetilde{H}$ comes into the problem. This has also been found by Bocquet *et al.* (2011, see also Wu *et al.*, 2011), who in an excellent article derive expressions for aggregation errors and also discuss applications of this formalism.

This sequence of approximations shows what is needed to be able to reduce the representation error related to a too low resolution of the model to a simple addition of a covariance matrix to the measurement error matrix. We reiterate that this covariance can be time dependent, and is generally smaller than that of the unconstrained high-resolution model variability.

We have to compare this treatment with that of Cohn (1997), who presented the first very careful analysis of representation errors. In his section 2.2 he decomposes the measurement equation as follows:

$$\begin{aligned} y &= H_{\mathrm{t}}(x_{\mathrm{t}}) + \epsilon_{\mathrm{m}} \\ &= H(x) + H_{\mathrm{t}}(x_{\mathrm{t}}) - H_{\mathrm{t}}(x) + H_{\mathrm{t}}(x) - H(x) + \epsilon_{\mathrm{m}} \\ &= H(x) + \epsilon_{\mathrm{r}} + \epsilon_{\mathrm{m}}\,, \end{aligned} \qquad (19)$$

in which x_{t} is the true state from which the actual observation was measured, H_{t} is the true observation operator, and $\epsilon_{\mathrm{r}} = H_{\mathrm{t}}(x_{\mathrm{t}}) - H_{\mathrm{t}}(x) + H_{\mathrm{t}}(x) - H(x)$. This error combination is the representation error and has contributions from the fact that the model state is a modelled version of the true state, and the observation operator H we actually use is not the true one. The second contribution will be dealt with in section 7, and the first contribution is what we have dealt with here.

Note that our approach is different from that of Janjic and Cohn (2006). They also assume Gaussian pdfs, but start from the unresolved scales. This then leads to two extra covariance operators: one for the unresolved scales, and one relating unresolved and resolved scales. To solve the full problem, they extend the state vector to include the unresolved scales, and them make approximations to this full system. By exploring a Kalman filter, they are able to estimate these covariances sequentially. Here we see that a full reference to the unresolved scales is not needed, only the relation between the unresolved and the resolved scales. However, we still have to specify $p(\widetilde{z}|x)$, and approximations will be needed in any large-scale problem. (Another difference is that Janjic and Cohn (2006), similarly to Cohn (1977), assume that the observation is a point measurement, in which case the unresolved variables are continuous, and covariance matrices become convolution operators working on model fields. We refrain from such a formulation, arguing that point measurement observations do not exist, and consequently assume observations are spatial averages.)

Finally, we reiterate that the representation errors might be correlated over time. Again we take the view that the unresolved scales will decorrelate between observation times, also because the conditioning will change because the mean state $\overline{H}x$ changes over time.

When the pdf of the unresolved scales is not Gaussian, we can approximate it with a standard kernel technique, e.g. Silverman (1986). Using a Gaussian as the kernel leads to the so-called Gaussian mixture density:

$$p(\widetilde{z}|x) = \sum_{i=1}^{M} c_i N(\mu_i, \Sigma_i), \qquad (20)$$

in which M is the number of mixture components. The $N(\mu_i, \Sigma)$ stands for the Gaussian density with μ_i the mean of the Gaussian pdf i, Σ_i is the covariance and c_i are normalization constants constrained by

$$\sum_{i=1}^{M} c_i = 1. \qquad (21)$$

There are standard techniques to estimate M, and the μ_i and Σ_i from the high-resolution runs, like the Expectation–Maximization technique, e.g. Bishop (2006), but since this

estimation problem is nonlinear it can be quite complex. Because Hx is assumed to be the mean over area i, the mean of the $\widetilde{z} = 0$ over this area, leading to

$$\bar{\widetilde{z}} = \int \widetilde{z}\, p(\widetilde{z}|x)\, \mathrm{d}\widetilde{z} = \sum_{i=1}^{M} c_i \mu_i = 0. \tag{22}$$

Using this expression in Eq. (15), assuming Gaussian distributed observations leads to

$$p(y|x) \propto \sum_i c_i N(\overline{H}x + \widetilde{H}\mu_i, C_\mathrm{m} + \widetilde{H} C_i \widetilde{H}^\mathrm{T}). \tag{23}$$

This formulation does require estimating the parameters of the Gaussian mixture model, but is very general because the Gaussian mixture model is quite general.

A potential problem with this expression in linearised data assimilation is that it is not in unimodal Gaussian form. If this expression is considered too complex to use in the existing assimilation scheme, we can approximate it by considering only its mean and covariance. Its mean is given by

$$\overline{y} = \int y p(y|x)\, \mathrm{d}y = \overline{H}x, \tag{24}$$

where we used $\sum_{i=1}^{M} c_i \mu_i = 0$, and the covariance becomes

$$\int (y-\overline{y})(y-\overline{y})^\mathrm{T} p(y|x)\, \mathrm{d}y = C_\mathrm{m} + \sum_i c_i \widetilde{H}(C_i + \mu_i \mu_i^\mathrm{T})\widetilde{H}^\mathrm{T}. \tag{25}$$

We can understand this result by realising that the covariance of the Gaussian mixture for $p(\widetilde{z}|x)$ is equal to

$$C_{\widetilde{z}|x} = \sum_i c_i (C_i + \mu_i \mu_i^\mathrm{T}), \tag{26}$$

so we find that the covariance for the observations y given the low-resolution model state x is the sum of the measurement errors of the observations C_m and a representation error due to the relation between the high- and low-resolution model.

With this Gaussian approximation of the representation error we find for the full likelihood:

$$p(y|x) \propto \exp\left[-\frac{1}{2}(y-\overline{H}x)^\mathrm{T} \left\{ C_\mathrm{m} + \sum_i c_i \widetilde{H}(C_i + \mu_i \mu_i^\mathrm{T})\widetilde{H}^\mathrm{T} \right\}^{-1} \times \left(y - \overline{H}x\right) \right], \tag{27}$$

where we identify the full observation error as $R = C_\mathrm{m} + \sum_i c_i \widetilde{H}(C_i + \mu_i \mu_i^\mathrm{T})\widetilde{H}^\mathrm{T}$. This equation sheds some light on the origin and on how to estimate the representation error when the high-resolution pdf given the low-resolution state is non-Gaussian, as will often be the case. Interestingly, we do not have to estimate the Gaussian mixture model parameters in this case because, as shown above, we use only the total covariance of that model, which can be estimated directly from the distribution of $\widetilde{z}$ given x.

6. Observations and state vector are of different types: assimilating retrievals

Another possibility is that the model does not have a direct equivalent of the observation, i.e. the process that governs the observation is not modelled at all. The previous section can be seen as a special case of what we discuss here. Again, one has to find the pdf that describes the relation between the observation and the model fields, and we can do that only by introducing other models.

Two possibilities exist to solve this problem. Firstly, the extra model can be incorporated directly in the original model, e.g. a radiative transfer model is built into the numerical weather prediction model. The advantage is that all statistics will be consistent, so the radiative transfer model will use the temperature and humidity field from the original model directly. A disadvantage is of course that the extra model can be quite expensive to run, which is not desirable when time is critical, as in numerical weather prediction.

A second option is to run the extra model off-line and assimilate the retrieved field into the original model. A potential problem is that, instead of assimilating the direct observations, one assimilates the extra model, which is not necessarily what one would like to do. An example is again radiance observations and an atmospheric circulation model that does not have a radiative transfer code. One could run the radiative transfer model off-line, and assimilate e.g the resulting temperature profiles. The problem is that the radiative transfer model with all its errors is assimilated, instead of the radiance. One of the potential errors is that the extra off-line model is run with a temperature and humidity field that is not the same as that of the actual model used in the assimilation.

As we will show below, there is actually no disadvantage in doing the assimilation off-line, and assimilating the retrieved variables at a later stage as long as proper care is taken. This has been discussed in great depth in Migliorini (2012) for the linear case, and below we will extend this to the nonlinear case, exploring Bayes' theorem.

6.1. General formulation

Let us denote a vector in the retrieval space as z, related to the original observations y via

$$y = H_\mathrm{ret}(z) + \epsilon\,. \tag{28}$$

As an example, y could be the radiation measurement from a satellite, and z an atmospheric temperature profile. H_ret is then the radiative transfer model.

We also need the relation between the retrieved value and the original model variables:

$$z = H_\mathrm{r}(x). \tag{29}$$

For numerical weather prediction, x could be the full state vector of a full atmospheric model, and z the temperature profile mentioned above. $H_\mathrm{r}(x)$ would in this case be the restriction of the full atmospheric model state to the temperature profile at the specific location connected to z. Clearly, one has to define H_ret and H_r such that $H(x) = H_\mathrm{ret}\{H_\mathrm{r}(x)\}$.

The retrieval problem can be derived from the original data assimilation problem exploring Bayes' theorem as follows:

$$p(x|y) = \frac{p(y|x)}{p(y)} p(x) = \frac{p(y|x)}{p(y)} \frac{p(x)}{p(z)} p(z), \tag{30}$$

in which we just multiplied the prior $p(x)$ with the prior of the retrieval $p(z)$, and divided by the same density. It is crucial to fully understand the notation: the pdf p(..) is defined by the argument, so $p(x)$ is a different pdf from $p(z)$. This equation can be rewritten as

$$p(x|y) = \left[\frac{p(y|x)}{p(y)} p(z) \right] \frac{p(x)}{p(z)}. \tag{31}$$

Because the direct interaction between the observation and the original model variable x is via z, so $p(y|x) = p\{y|H_r(x)\} = p(y|z)$,

we can interpret the first two factors in brackets in this equation as the retrieval process. To find the posterior for the original data-assimilation problem we have to multiply the retrieval pdf by $p(x)/p(z)$, so

$$p(x|y) = \frac{p(z|y)}{p(z)} p(x). \tag{32}$$

A few remarks can be made about the process outlined above. Firstly, we see that the retrieval is in fact a full pdf, which allows us to incorporate the nonlinear case quite naturally. Secondly, this retrieval can be 'assimilated' in the original model by a simple multiplication with the original prior divided by the prior of the retrieval. Thirdly, since $z = H_r(x)$, the denominator depends on x, contrary to the direct assimilation of the observations into the original model. To eliminate this dependency, one might argue to store the retrieval as $p(z|y)/p(z)$, and assimilating that pdf into the original model reduces to a simple multiplication by $p(x)$. However, note that $p(z|y)/p(z) = p(y|z)/p(y)$, which, as function of z is not unique because $H_{\rm ret}$ can typically not be inverted. This means that either one first calculates $p(x)/p(z)$ and multiplies that with the retrieval pdf, or one combines retrieval, $p(x)$ and $p(z)$ in one process. Below we will evaluate this general formalism in linear Gaussian and nonlinear context.

Finally, it is possible that a representation error is present between the observations y and the retrieval variable z. In that case one can follow the methods explained in sections 3–5, or 7, dependent on the issue. It is also possible that the retrieval z represents reality differently from the model state x. In that case we can write the retrieval as

$$z = H_r(x) + \widetilde{z}, \tag{33}$$

and proceed as

$$\begin{aligned} p(x|y) &= p(x) \int \frac{p(y|z)}{p(y)} p(\widetilde{z}|x)\, \mathrm{d}\widetilde{z} \\ &= p(x) \int \frac{p(y|z)p(z)}{p(y)} \frac{p(\widetilde{z}|x)}{p(z)}\, \mathrm{d}\widetilde{z}, \end{aligned} \tag{34}$$

in which we recognise the retrieval as the first factor in the integral. However, the change from standard retrieval case is that we have to integrate over all possible values of the high-resolution retrieval, as in sections 3 and 4. So one has to build the pdf of the high-resolution retrieval variable given the coarse-resolution dynamical model variable x.

6.2. The linear Gaussian case

For the linear Gaussian case we assume all pdfs and the likelihood are Gaussian distributed. In particular the prior for the retrieval is $p(z) = N(z_b, B_{ret})$. This leads to a retrieval

$$\begin{aligned} p(z|y) &= \frac{p(y|z)}{p(y)} p(z) \\ &\propto \exp\left\{ -\frac{1}{2} (z - z_{\rm ret})^{\rm T} P_{\rm ret}^{-1} (z - z_{\rm ret}) \right\} \end{aligned} \tag{35}$$

in which

$$z_{\rm ret} = z_{\rm b} + K_{\rm ret} \left(y - H_{\rm ret} z_{\rm b} \right) \tag{36}$$

with retrieval gain

$$K_{\rm ret} = B_{\rm ret} H_{\rm ret}^{\rm T} \left(H_{\rm ret} B_{\rm ret} H_{\rm ret}^{\rm T} + R \right)^{-1} \tag{37}$$

and posterior covariance

$$P_{\rm ret} = \left(1 - K_{\rm ret} H_{\rm ret} \right) B_{\rm ret}, \tag{38}$$

which is the standard solution. As mentioned above, to 'assimilate' this retrieval into the original model it has to be multiplied by $p(x)/p(z)$ leading to

$$\begin{aligned} \frac{p(z|y)}{p(z)} p(x) \propto \exp\Bigg[&-\frac{1}{2} (z - z_{\rm ret})^{\rm T} P_{\rm ret}^{-1} (z - z_{\rm ret}) \\ &+ \frac{1}{2} (z - z_{\rm b})^{\rm T} B_{\rm ret}^{-1} (z - z_{\rm b}) \\ &- \frac{1}{2} (x - x_{\rm b})^{\rm T} B^{-1} (x - x_{\rm b}) \Bigg]. \end{aligned} \tag{39}$$

Using $z = H_r x$ and completing the squares in x we find

$$\frac{p(z|y)}{p(z)} p(x) \propto \exp\left[-\frac{1}{2} (x - x_{\rm a})^{\rm T} P^{-1} (x - x_{\rm a}) \right] \tag{40}$$

in which $x_{\rm a}$ is found as

$$x_{\rm a} = P B^{-1} x_{\rm b} + P H_{\rm r}^{\rm T} \left(P_{\rm ret}^{-1} z_{\rm ret} - B_{\rm ret}^{-1} z_{\rm b} \right) H_{\rm r} \tag{41}$$

with

$$P^{-1} = B^{-1} + H_{\rm r}^{\rm T} \left(P_{\rm ret}^{-1} - B_{\rm ret}^{-1} \right) H_{\rm r}\,. \tag{42}$$

These are the equations that assimilate the retrieval into the original model. We have to store not only the retrieval $z_{\rm ret}$ and its error covariance $P_{\rm ret}$, but also retain the retrieval prior $z_{\rm b}$ and its covariance $B_{\rm ret}$.

Note that we cannot write the solution in usual covariance form, mainly because there is no equivalent for y in terms of $z_{\rm b}$ and $z_{\rm ret}$. This is directly related to the fact the the information matrix of $p(z|y)/p(z)$ is given by

$$A = P_{\rm ret}^{-1} - B_{\rm ret}^{-1} = H_{\rm ret}^{\rm T} R^{-1} H_{\rm ret}, \tag{43}$$

and the latter expression cannot be inverted if $H_{\rm ret}$ is not full rank.

To make the connection to present-day data assimilation methods used in numerical weather forecasting, variational algorithms would minimise minus the logarithm of the posterior, and so minimise

$$\begin{aligned} J(x) = &\frac{1}{2} (H_{\rm r} x - z_{\rm ret})^{\rm T} P_{\rm ret}^{-1} (H_{\rm r} x - z_{\rm ret}) \\ &- \frac{1}{2} (H_{\rm r} x - z_{\rm b})^{\rm T} B_{\rm ret}^{-1} (H_{\rm r} x - z_{\rm b}) \\ &+ \frac{1}{2} (x - x_{\rm b})^{\rm T} B^{-1} (x - x_{\rm b}). \end{aligned} \tag{44}$$

No specific problem arise here since the gradient can be calculated directly.

The application of ensemble Kalman filters is less straightforward as they are typically derived from the covariance form. Below we explain how one can use the Ensemble Transform Kalman Filter (ETKF). We can write:

$$\begin{aligned} P &= \left\{ B^{-1} + H_{\rm r}^{\rm T} \left(P_{\rm ret}^{-1} - B_{\rm ret}^{-1} \right) H_{\rm r} \right\}^{-1} \\ &= \left[B^{-1/2} \left\{ 1 + B^{1/2} H_{\rm r}^{\rm T} \left(P_{\rm ret}^{-1} - B_{\rm ret}^{-1} \right) H_{\rm r} B^{1/2} \right\} B^{-1/2} \right]^{-1} \\ &= B^{1/2} \left\{ 1 + B^{1/2} H_{\rm r}^{\rm T} \left(P_{\rm ret}^{-1} - B_{\rm ret}^{-1} \right) H_{\rm r} B^{1/2} \right\} B^{1/2} \\ &= X \left\{ 1 + (H_{\rm r} X)^{\rm T} \left(P_{\rm ret}^{-1} - B_{\rm ret}^{-1} \right) H_{\rm r} X \right\} X^{\rm T} \\ &= X T T^{\rm T} X^{\rm T}, \end{aligned} \tag{45}$$

in which we used the standard notation $B = XX^{\rm T}$ in which X is the ensemble perturbation matrix defined as $X = 1/\sqrt{N-1}(x_1 - x_{\rm b}, \ldots, x_N - x_{\rm b})$ for an ensemble of size N. In an ETKF one used the matrix T to update the ensemble perturbations as $X^a = XT$, and the above gives an expression for this T. As in the ETKF we

can relate the terms in the transform matrix T to the ensemble, exploring SVDs etc. Details can be found in the ETKF literature (e.g. Bishop *et al.*, 2001). The ensemble mean is found from Eq. (41) as

$$x_a = (1-KH_r)x_b + XTT^T(H_rX)^T\left(P_{ret}^{-1}z_{ret} - B_{ret}^{-1}z_b\right)H_r, \quad (46)$$

with

$$K = PH^TA = XTT^T(H_rX)^T(P_{ret}^{-1} - B_{ret}^{-1}). \quad (47)$$

This closes the linear retrieval case, and elements of this derivation can be found in e.g. Migliorini (2012), who discusses this issue in great detail. Specifically he mentions two conditions that are needed for the equivalence of assimilating observations directly and assimilating retrievals. The first one is that a linear approximation of the observation operator is adequate, essentially making his retrieval problem linear. The second one is that the prior information covariance used in the retrieval should not be taken too narrow to avoid losing observation information. This latter point is not discussed here, but is obviously a very important point for practical applications.

6.3. *The nonlinear case*

As mentioned above, the retrieval is in fact a full pdf. Because of the direct connection to Bayes' theorem, we know immediately how to use the retrieval in the nonlinear case. Several nonlinear data assimilation methods exist and we will discuss particle filtering here since these can be applied in high-dimensional systems (e.g. van Leeuwen, 2011; Ades and van Leeuwen, 2014).

We start again from our general expression for assimilation via a retrieval:

$$p(x|y) = \frac{p(y|x)}{p(y)}\frac{p(x)}{p(z)}p(z). \quad (48)$$

In importance sampling one considers samples from the prior of the retrieval $p(z)$ and incorporates the other terms as additional weights on these samples. So we start from

$$p(z) = \sum_{i=1}^{N_z}\delta(z - z_i), \quad (49)$$

and plug this into Bayes for the retrieval, leading to

$$\begin{aligned} p(z|y) &= \frac{p(y|z)}{p(y)}p(z) = \sum\frac{1}{N}\frac{p(y|z_i)}{p(y)}\delta(z - z_i) \\ &= \sum w_i\delta(z - z_i), \end{aligned} \quad (50)$$

so the retrieved particles are now weighted with

$$w_i = \frac{1}{N}\frac{p(y|z_i)}{p(y)} \propto p(y|z_i). \quad (51)$$

Returning to the full assimilation, we find

$$p(x|y) = \frac{p(y|x)}{p(y)}\frac{p(x)}{p(z)}p(z) = \sum w_i\frac{p\,(x|z_i)}{p(z_i)}\delta(z - z_i). \quad (52)$$

The meaning of the extra weights $p(x|z_i)/p(z_i)$ is as follows. First $p(z_i)$ is the probability of particle z_i in the retrieval prior $p(z)$. For example, if the retrieval prior is a Gaussian, this can be evaluated as

$$p(z_i) \propto \exp\left[-\frac{1}{2}(z_i - z_b)^TB_{ret}^{-1}(z_i - z_b)\right], \quad (53)$$

so just a number we can calculate directly. Note that we do not need to worry about the normalisation constant of the Gaussian, as it is the same for each particle and is taken care of automatically when we normalise the sum of the weights to one. The other factor $p(x|z_i)$ is a bit more problematic. First we have to realise that, when z_i is given, only part of the full state vector x is known, namely the part $H_r(x)$ (in which $H_r(..)$ can be a nonlinear operator). So, we have to evaluate this term as follows:

$$p(x|z_i) = p\left\{x|H_r(x) = z_i\right\}. \quad (54)$$

If we could know $p(x)$ for every x, we would be able to calculate this, at least in principle. However, this is typically not the case in nonlinear data assimilation.

To proceed, let us assume the original model uses a particle filter. In the simplest particle filter, $p(x)$ is represented by a sum of delta functions as

$$p(x) = \sum_{j=1}^{N_x}\frac{1}{N}\delta(x - x_j), \quad (55)$$

where we assumed equal weight particles, but including weighted particles is straightforward. The issue is now that typically $H_r(x_j) \neq z_i$ for every i, j combination: the retrieval prior and the prior of the original model are not generated simultaneously. There are several ways to solve this problem, and some are valid only approximately. One could interpolate in state space the posterior retrieval pdf from the weighted prior retrieval samples z_i. Since the dimension of the z_i is much smaller than that of the x_j, this should be easier than interpolating in x space.

Another option is to kernel dress the particles, i.e. assume each particle is the mean of a known function, say a Gaussian, with pre-specified width (covariance). Or, similarly, one could fit a Gaussian mixture model. This can be done on either the posterior retrieval, or on the prior of the original model, or both. In that case either $p(z_j|y)$, $p(x_i)$, or both can be calculated directly from the interpolated pdfs.

Typically, however, a more sophisticated particle filter will be used, which explores a proposal density (e.g. van Leeuwen, 2009). In short, instead of using the model equations from the original model to arrive at the observation time, one can use modified equations that include information about the observations in the form of e.g. a relaxation term, or even a full 4D-Var. The fact that one uses a modified equation can be compensated for by changing the relative weights of the particles. In our case one could change the model equations also to include information on the location of the z_i. In its simplest form this could be implemented as a relaxation term that relaxes each x_j to one of the z_i, as well as to the rest of the observations. If needed, one can add steps as in the Equivalent-Weights Particle Filter to ensure that the particles of the original model end up close enough to the posterior retrieval particles (e.g. van Leeuwen, 2010; van Leeuwen, 2011; Ades and van Leeuwen, 2013). Clearly several possibilities are open here, and it is unclear what the best technique is. We will discuss an example below.

If we consider present-day large-scale data assimilation systems, they are typically either linear or based on linearisations. In that case it can be useful to approximate the retrieval pdf by a Gaussian, while still using a nonlinear retrieval. The idea is that a linear retrieval will be biased, and the retrieved covariance will be biased too for a nonlinear retrieval process. However, if a fully nonlinear data assimilation method for the retrieval is used, the first two moments of the retrieved pdf are unbiased.

We have found an expression for the assimilation of nonlinear retrievals, showing that this is possible without loss of information from the original observations. Practically, one either has to interpolate or use proper proposal densities to make this a valid statement.

6.4. *Examples*

We will now discuss two examples to illustrate the use of this equation, one analytical and one exploring particle filters. In the first example we assume that observation and model state are related by

$$y = |H_r x| + \epsilon\,, \tag{56}$$

in which H_r is the projection of the full model state to a subspace of the full model state, e.g. a model grid point, or a small model area, and ϵ is a random variable drawn by the measurement process from $N_\epsilon(0, R)$. Note the subscript on N to denote the active variable. The retrieved variable is defined as the model state subspace $z = H_r x$, so

$$y = |z| + \epsilon\,. \tag{57}$$

The prior for the retrieval is taken as a pdf given by $N_z(z_b, B_{ret})$. We then find for the retrieval from Bayes' theorem:

$$p(z|y) \propto \begin{cases} N_z(y, R) & N_z(z_b, B_{ret}) \quad \text{if } z \geq 0, \\ N_z(-y, R) & N_z(z_b, B_{ret}) \quad \text{if } z < 0, \end{cases}$$

which can be evaluated as:

$$\propto \begin{cases} N_z\{z_b + K_{ret}(y - z_b) & , P_{ret}\} \quad \text{if } z \geq 0, \\ N_z\{z_b + K_{ret}(-y - z_b), & P_{ret}\} \quad \text{if } z < 0, \end{cases}$$

in which $K_{ret} = B_{ret} H_{ret}^T (H_{ret} B_{ret} H_{ret}^T + R)^{-1}$ and $P_{ret} = (1 - K_{ret} H_{ret}) B_{ret}$ (compare with the linear case). This is the retrieved pdf, which we now assimilate into the full model. Note that the retrieved pdf is bimodal with Gaussian distributions around the two modes $z_b + K_{ret}(y - z_b)$ and $z_b + K_{ret}(-y - z_b)$.

How we proceed from here depends on the prior of the original model. Assuming that prior is Gaussian $N_x(x_b, B)$ (or a Gaussian mixture, the extension is straightforward), we find for the posterior pdf (see the linear case):

$$p(x|y) \propto \begin{cases} N_x(x_1, P) & \text{if } H_r x \geq 0, \\ N_x(x_2, P) & \text{if } H_r x < 0, \end{cases}$$

in which $x_1 = PB^{-1}x_b + PH_r^T \left(P_{ret}^{-1} z_{ret} - B_{ret}^{-1} z_b\right) H_r$, with $z_{ret} = z_b + K_{ret}(y - z_b)$
and $x_2 = -x_1\{z_{ret} = z_b + K_{ret}(-y - z_b)\}$.
This closes this example.

We close the discussion with a numerical example using particle filters. Assume that the observation operator reads as

$$y = x^2 + \epsilon\,, \tag{58}$$

in which $\epsilon \sim N(0, R)$, with variance $R = 0.01$. As retrieval prior we use a Gaussian as

$$p(z) = N(0, B_{ret}) \tag{59}$$

with $B_{ret} = 0.4$.

Figures 2 and 3 show the retrieval prior and retrieval posterior pdfs using an ensemble of 50 particles. The effective ensemble size, defined as $1/w_i^2$ was 16. Experiments with smaller numbers show similar results but with increased sampling noise. We assume the two-dimensional prior of the original model 1 time step before the time of the observation (or retrieval) is given by a Gaussian

$$p(x) = N(0, B) \tag{60}$$

to allow for an exact solution for comparison, with, for simplicity, $B = \text{diag}(0.01)$.

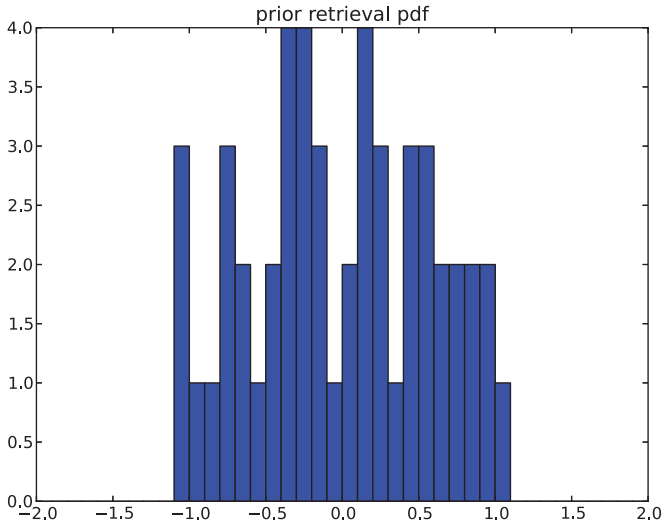

Figure 2. Retrieval prior pdf using an ensemble of 50 particles against value of x.

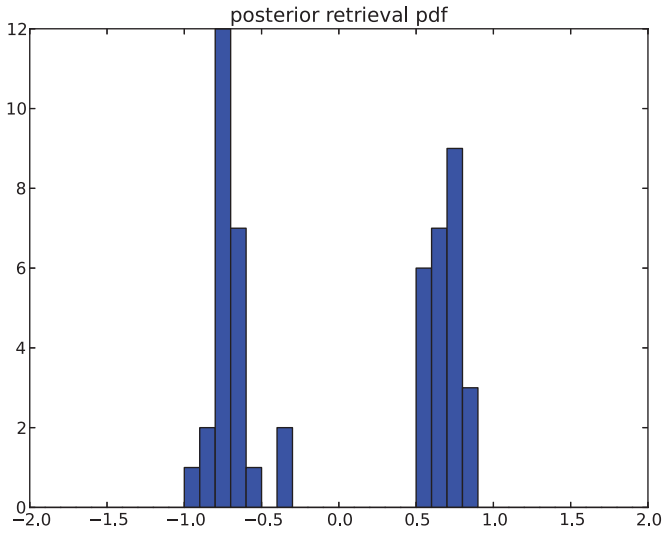

Figure 3. As Figure 2, but showing the retrieval posterior pdf against value of x.

The model evolution equation is taken as the identity

$$x^{n+1} = x^n + \beta^n, \tag{61}$$

in which the model error is $\beta \sim N(0, Q)$, with the model error covariance Q given by:

$$Q = \begin{pmatrix} 0.1 & 0.05 \\ 0.05 & 0.2 \end{pmatrix}. \tag{62}$$

The observation operator between model state and retrieval state is now linear as all nonlinearity is taken up by the retrieval, so $H_r(x) = x^1$, the first element of x.

To ensure convergence to the posterior retrieval particles while retaining non-degenerate particles, we use the Equivalent-Weights Particle Filter scheme (e.g. Ades and van Leeuwen, 2013). This consists of employing a modified model equation

$$x_j^{n+1} = x_j^n + \widehat{\beta}_j^n + \alpha_j T\{z_i - H_r(x_j)\}, \tag{63}$$

in which $T = QH_r^T(H_r Q H_r^T + \widehat{R})^{-1}$ and $\widehat{R}$ is small. We choose $\widehat{R} = 0.00001$, which will ensure that the $H_r(x_j)$ part of x_j is forced to be almost equal to z_i. The random term $\widehat{\beta}_j^n$ is chosen from a mixture density with small amplitude, (Ades and van Leeuwen, 2013, 2014, give details). We retain 80% of

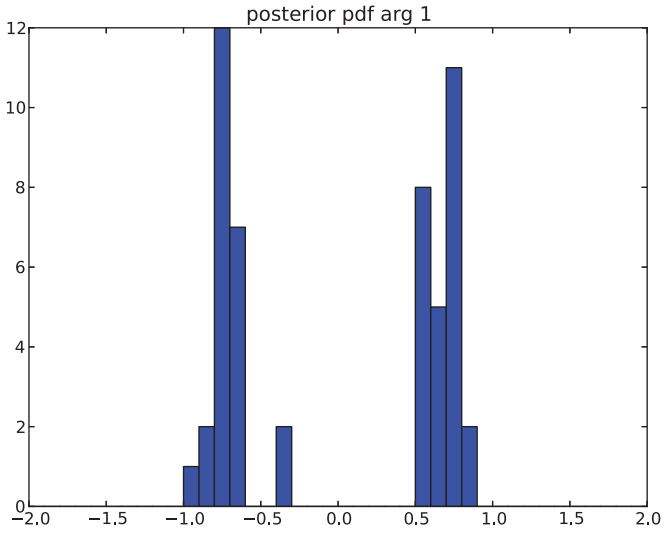

Figure 4. Model posterior pdf for argument 1 after the assimilation of retrieval pdf against value of x.

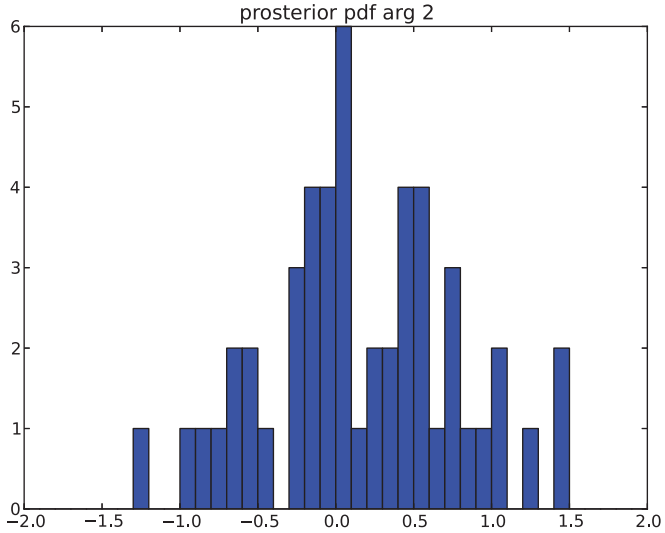

Figure 6. Model posterior pdf for argument 2 after assimilation of retrieval pdf against value of x.

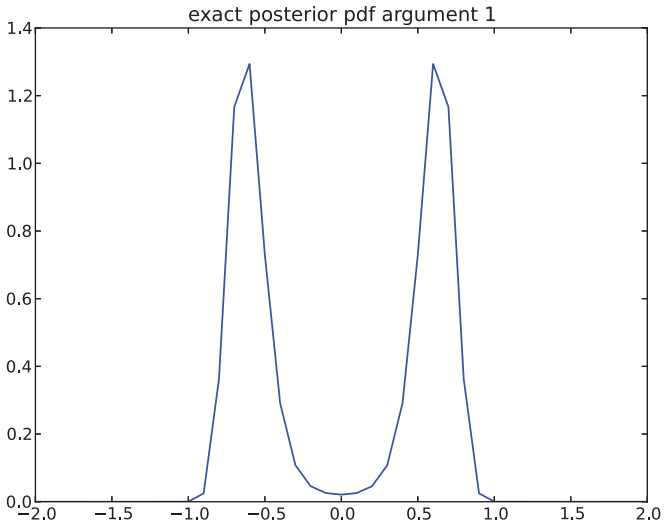

Figure 5. Exact posterior pdf for direct assimilation of observation, argument 1 against value of x.

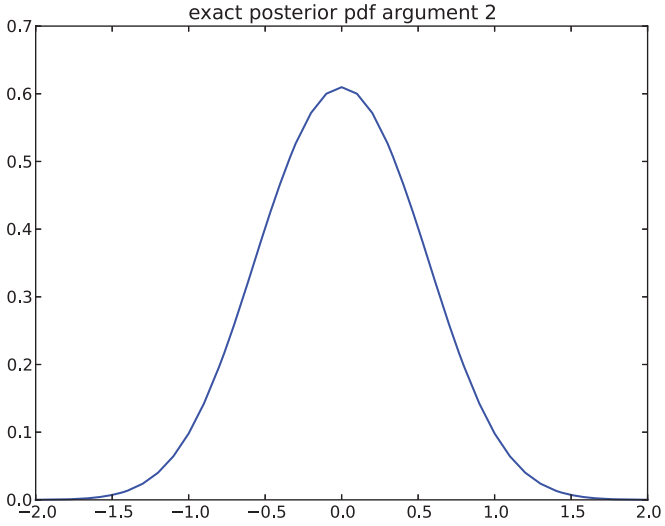

Figure 7. Exact posterior pdf for direct assimilation of observation, argument 2 against value of x.

the particles but little difference was seen in the range 70–100%. The scheme chooses α_j such that the weights of all particles are equal, taking into account the weights accumulated over previous time steps. These accumulated weights are $1/p(z_i)$ in this case (Eq. (52)).

Figure 4 shows the posterior pdf for $x^1 = H_r(x)$ and Figure 5 shows the exact solution. The two are very close (apart from sampling noise), with the left peak having 24 members and the right peak 26 members, showing that assimilating the retrieval works. The effective ensemble size was 40, as expected since 80% of the particles were retained.

The second element of x_j is also changed due to the non-zero covariances in Q, leading to the posterior marginal for that element depicted in Figure 6, with exact solution in Figure 7. This closes the example on how to assimilate nonlinear retrievals using a particle filter into the original model represented by a set of particles.

7. The likelihood with non-exact observation operator

When the relation between model and observations is not known exactly, an extra step has to be taken into account, as identified by e.g. Cohn (1977). Note that this problem is different from that in section 3. There, model and observations represented the truth differently. Here we assume that the representation is similar, but we do not know exactly how the two are related. For instance, the observation operator used to connect model state to observation state has parameters that are not well known. The likelihood will depend on the observation operator used, and the pdf representing the uncertainty of the observation operator has to be taken into account. This can be done by bringing in the observation operator as an extra random variable. The likelihood can now be written as

$$p(y|x) = \int p(y, H|x)\, \mathrm{d}H\,. \tag{64}$$

This equation can be evaluated as

$$p(y|x) = \int p(y, H|x)\, \mathrm{d}H = \int p(y|H, x)\, p(H|x)\, \mathrm{d}H\,. \tag{65}$$

The first factor $p(y|H, x)$ can be evaluated in the usual way, as discussed in the previous sections, including the representation error in case our modelled state is a low-resolution version of the true state. The second factor $p(H|x)$ is the pdf of the observation operator given state x. If H is linear it will not depend on x so $p(H|x) = p(H)$. This pdf has to be known to be able to solve the data assimilation problem.

In general little can be said about this solution. To make some progress, let us now concentrate on the special case that both $H|x$ and $y|H, x$ are Gaussian distributed. We then have

$$p(y|x) \propto \int \exp\left[-\frac{1}{2}\left\{y - H(x)\right\} C_{\mathrm{m}}^{-1} \left\{y - H(x)\right\}^{\mathrm{T}} - \frac{1}{2}\{H(x) - h(x)\} S^{-1} \{H(x) - h(x)\}^{\mathrm{T}}\right] \mathrm{d}H\,, \quad (66)$$

in which h is our best guess of the observation operator, and S its error covariance. When H is linear we can reduce this further to

$$p(y|x) \propto \exp\left[-\frac{1}{2}\left\{y - h(x)\right\} (C_{\mathrm{m}} + S)^{-1} \left\{y - h(x)\right\}^{\mathrm{T}}\right]. \quad (67)$$

So, in the linear Gaussian case we can just add the error covariance of the observation operator model to the error covariance of the measurement noise. Note that this equation holds also when $h(x)$ is nonlinear and the uncertainty in the observation operator is in the linear regime. This is the case discussed in Lorenc (1986) and Bocquet *et al.* (2011).

Let us finally bring the results from sections 3–5 and 7 together. If the model is less realistic than the observation, and the observation operator is not known exactly, the likelihood becomes, using z as the high-resolution model variable,

$$p(y|x) = \int p(y|x, z)\, p(z|x)\, \mathrm{d}z = \int p(y|z)\, p(z|x)\, \mathrm{d}z = \int p(y|H, z)\, p(H|z)\, p(z|x)\, \mathrm{d}H \mathrm{d}z\,. \quad (68)$$

When all pdfs are assumed to be Gaussian and H linear, we can evaluate this as

$$p(y|x) \propto \exp\left[-\frac{1}{2}\left\{y - \widetilde{h}(x)\right\}^{\mathrm{T}} \left(C_{\mathrm{m}} + S + \widetilde{\widetilde{C}}\widetilde{h}^{\mathrm{T}}\right)^{-1} \left\{y - \widetilde{h}(x)\right\}\right], \quad (69)$$

in which $\widetilde{h}$ is the mean of the observation operator in the low-resolution model, with error covariance S, and $\widetilde{C}$ is the covariance matrix of the high-resolution fields, given the low-resolution value x.

Finally, other approximations than the pure Gaussian can be used to describe representation errors of this kind. The simplest analytically is the Gaussian mixture model explored earlier, but other methods can be explored also.

8. Conclusions

This article discusses the formal treatment of the so-called representation error using Bayes' theorem. It is shown how the likelihood can be interpreted and calculated when the model and the observations represent different phenomena. The representation error issue in observation space is discussed, and we showed how the Gaussian assumption leads to the traditional formulation of the representation error. It is stressed that the part of the likelihood related to representation mismatch is dependent on the model state in general. For the Gaussian case, this means that the representation error covariance is model state dependent, and typically smaller than the subgrid-scale covariance.

We also discuss the idea of averaging observations (superobbing) to avoid the problem of the representation error and show that this does not work: one cannot get rid of the representation error, and when the observations are correlated the relative influence of the representation error increases when more observations are used in the averaging.

When the representation error is solved for via model simulation, we find again that the representation mismatch depends on the model state, and similar caution as above should apply here. Several strategies are discussed to solve the problem using high-resolution model simulations.

The off-line versus online retrieval problem is discussed and using Bayes' theorem a new simple proof is given of the equivalence of assimilating the original observations and assimilating their linear retrievals. This proof is extended to nonlinear retrievals, again using Bayes' theorem.

Furthermore, we discuss how to treat uncertainty in the observation operator model. In the most general case, one has to solve a convolution of the likelihood given a certain observation operator model and the pdf of the observation operator model itself. It is shown that, only when both are Gaussian and H are linear (or linearisable around a mean observation operator model), the error covariance in the likelihood becomes the sum of the error covariances of the likelihood given an observation operator model and the error covariance of the observation operator model.

While no real-world applications have been discussed, it is hoped that this article clarifies some of the issues around representation errors.

Acknowledgements

I thank the National Centre for Earth Observation (NCEO) and the National Environmental Research Council (NERC) for support via several grants. Also two anonymous reviewers of an earlier version are thanked for critical comments which improved the article substantially.

Appendix

Equivalence of Eq. (13) and (14)

We will prove that assimilating a set of subgrid observations $(y_1, ..., y_m)^{\mathrm{T}}$ with identical variances $\sigma_m^2 + \sigma_r^2$ and correlations ρ leads to the same analysis as assimilating their average $\overline{y}$. We assume that the model equivalent of the observation is given by Hx, in which H can be a model grid box, or a spatial average over a few model grid boxes.

For the update of the average we have

$$x_{av} = x + (B^{-1} + H^{\mathrm{T}}\sigma_{\overline{y}}^{-2}H)^{-1} H^{\mathrm{T}}\sigma_{\overline{y}}^{-2}(\overline{y} - Hx). \quad (A1)$$

The variance of the observation average is given by

$$\sigma_{\overline{y}}^2 = \sigma_{\overline{y}}^2 = \frac{\sigma_m^2 + \sigma_r^2}{m} + \frac{m-1}{m}\rho\sigma_r^2. \quad (A2)$$

To show how to work out these term for each specific case, we use the following example. Assume that H is an average over five model grid points and we arrange the ordering in the state vector such that these five grid points are positioned next to each other. Then H will be a matrix of size $1 \times n$ with value $1/5$ at the five grid points and zeros elsewhere:

$$H = (0, ..., 0, 1/5, 1/5, 1/5, 1/5, 1/5, 0, ..., 0). \quad (A3)$$

Clearly $H^{\mathrm{T}}\sigma_{\overline{y}}^{-2}H$ will be an $n \times n$ matrix with values $1/5 \times 1/5 \times \sigma_{\overline{y}}^2 = 1/(25\sigma_{\overline{y}}^2)$ at the entries (i, j) in which i and j relate to the five grid points:

$$H^{\mathrm{T}}\sigma_{\overline{y}}^{-2}H = \begin{pmatrix} 0, \ldots & 0 & 0, \ldots \\ 0, \ldots & 1/(25\sigma_{\overline{y}}^2) & 0, \ldots \\ 0, \ldots & 0 & 0, \ldots \end{pmatrix}.$$

The other observation-specific term in the update is $H^{\mathrm{T}}\sigma_{\overline{y}}^{-2}(\overline{y} - Hx)$. Note that because we have just one observation,

$H^{\mathrm{T}}\sigma_{\overline{y}}^{-2}$ will be an $n \times 1$ matrix with values $1/(5\sigma_{\overline{y}}^2)$ at the five grid point positions, and zero elsewhere. This finally leads to $H^{\mathrm{T}}\sigma_{\overline{y}}^{-2}(\overline{y} - Hx)$ being a $n \times 1$ matrix with entries $(\overline{y} - Hx)/(5\sigma_{\overline{y}}^2)$ at the five grid points, and zeros elsewhere.

Let us now look at the more complicated calculation when we assimilate all observations at the same time. The update for this case is given by:

$$x_{\mathrm{d}} = x + (B^{-1} + H_{\mathrm{d}}^{\mathrm{T}} R_{\mathrm{d}}^{-1} H_{\mathrm{d}})^{-1} H_{\mathrm{d}}^{\mathrm{T}} R_{\mathrm{d}}^{-1}(y - H_{\mathrm{d}}x), \tag{A4}$$

in which $y = (y_1, ..., y_m)^{\mathrm{T}}$, and H_{d} is an $m \times n$ matrix with zeros everywhere except in the columns related to the five grid points:

$$H_{\mathrm{d}} = \begin{pmatrix} 0, \ldots 0 & 1, 1, 1, 1, 1 & 0, \ldots \\ \vdots & \vdots & \vdots \\ 0, \ldots 0 & 1, 1, 1, 1, 1 & 0, \ldots \end{pmatrix}.$$

The observation-error covariance R_{d} is a matrix with $\sigma^2 + \sigma_{\mathrm{r}}^2$ on the diagonal and $\rho\sigma_{\mathrm{r}}^2$ at all other entries:

$$R_{\mathrm{d}} = \begin{pmatrix} \sigma^2 + \sigma_{\mathrm{r}}^2 & \rho\sigma_{\mathrm{r}}^2 & \rho\sigma_{\mathrm{r}}^2 \\ \rho\sigma_{\mathrm{r}}^2 & \sigma^2 + \sigma_{\mathrm{r}}^2 & \rho\sigma_{\mathrm{r}}^2 \\ \rho\sigma_{\mathrm{r}}^2 & \rho\sigma_{\mathrm{r}}^2 & \sigma^2 + \sigma_{\mathrm{r}}^2 \end{pmatrix}.$$

The inverse of R_{d} is easily found as a $m \times m$ matrix with

$$\frac{\sigma^2 + \sigma_{\mathrm{r}}^2 + (m-2)\rho\sigma_{\mathrm{r}}^2}{(\sigma^2 + (1-\rho)\sigma_{\mathrm{r}}^2)(\sigma^2 + \sigma_{\mathrm{r}}^2 + (m-1)\rho\sigma_{\mathrm{r}}^2)} \tag{A5}$$

on the diagonal and

$$\frac{-\rho\sigma_{\mathrm{r}}^2}{(\sigma^2 + (1-\rho)\sigma_{\mathrm{r}}^2)(\sigma^2 + \sigma_{\mathrm{r}}^2 + (m-1)\rho\sigma_{\mathrm{r}}^2)} \tag{A6}$$

in all other entries. (This can be checked simply by multiplying R_{d} with this matrix.) It is easy to see that $R_{\mathrm{d}}^{-1}H$ will be a $m \times n$ matrix with entries $1/(5m\sigma_{\overline{y}}^2)$ at the columns related to the five grid points:

$$R_{\mathrm{d}}^{-1}H = \frac{1}{5m\sigma_{\overline{y}}^2} \begin{pmatrix} 0, \ldots 0 & 1, 1, 1, 1, 1 & 0, \ldots \\ \vdots & \vdots & \vdots \\ 0, \ldots 0 & 1, 1, 1, 1, 1 & 0, \ldots \end{pmatrix}.$$

Here we use the fact that the diagonal of R_{d}^{-1} plus $m-1$ times its off-diagonal elements is equal to $1/\{\sigma^2 + \sigma_{\mathrm{r}}^2 + (m-1)\rho\sigma_{\mathrm{r}}^2\}$ which is equal to $1/(m\sigma_{\overline{y}}^2)$.

Similarly, multiplying this matrix by H^{T} leads to a $n \times n$ matrix with entries $1/(25\sigma_{\overline{y}}^2)$ at those positions (i, j) for which i and j correspond to the five grid points. So we see directly that the $H^{\mathrm{T}}R^{-1}H$ combination is the same for both assimilation methods.

We now show the same for the $H^{\mathrm{T}}R_{\mathrm{d}}^{-1}(y - Hx)$ factor. Clearly $H^{\mathrm{T}}R_r^{-1}$ is a $n \times m$ matrix with non-zero entries $1/(5m\sigma_{\overline{y}}^2)$ at the rows related to the five grid points. Multiplying this matrix with the innovation vector leads directly to an $n \times 1$ matrix with entries $(\overline{y} - Hx)/(5\sigma_{\overline{y}}^2)$ at the five grid points, and zeros elsewhere. Again, we find the same matrix as for the case in which we assimilated the average observation. This concludes the proof.

References

Ades M, van Leeuwen PJ. 2013. An exploration of the equivalent-weights particle filter. *Q. J. R. Meteorol. Soc.* **139**: 820–840, doi: 10.1002/qj.1995.

Ades M, van Leeuwen PJ. 2014. The equivalent-weights particle filter in a high-dimensional system. *Q. J. R. Meteorol. Soc.*, doi: 10.1002/qj.2370.

Anderson JL, Wyman B, Zhang S, Hoar T. 2005. Assimilation of surface pressure observations using an ensemble filter in an idealized global atmospheric prediction system. *J. Atmos. Sci.* **62**: 2921–2938.

Bishop CM. 2006. *Pattern Regocnition and Machine Learning*. Springer: New York, NY.

Bishop CH, Etherton B, Majumdar SJ. 2001. Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Mon. Weather Rev.* **129**: 420–436.

Bocquet M, Wu L, Chevallier F. 2011. Bayesian control space for optimal assimilation of observations, I: Consistent multiscale formalism. *Q. J. R. Meteorol. Soc.* **137**: 1340–1356.

Cohn SE. 1997. An introduction to estimation theory. *J. Meteorol. Soc. Jpn.* **75**: 257–288.

Daley R. 1993. Estimating observation error statistics for atmospheric data assimilation. *Ann. Geophys.* **11**: 634–647.

Derber J, Rosati A. 1989. A global oceanic data assimilation system. *J. Phys. Oceanogr.* **19**: 1333–1347.

Janjic T, Cohn SE. 2006. Treatment of observation error due to unresolved scales in atmospheric data assimilation. *Mon. Weather Rev.* **134**: 2900–2915.

Köhl A, Dommenget D, Ueyoshi K, Stammer D. 2007. 'The global ECCO 1952–2001 ocean synthesis'. Report Series 40. ECCO: Cambridge, MA.

Leeuwenburgh O. 2007. Validation of an EnKF system for OGCM initialization assimilating temperature, salinity, and surface height measurements. *Mon. Weather Rev.* **135**: 125–139.

Liu Z-Q, Rabier F. 2002. The interaction between model resolution, observation resolution and observation density in data assimilation: A one-dimensional study. *Q. J. R. Meteorol. Soc.* **128**: 1367–1386.

Lorenc AC. 1986. Analysis methods for numerical weather prediction. *Q. J. R. Meteorol. Soc.* **112**: 1177–1194.

Migliorini S. 2012. On the equivalence between radiance and retrieval assimilation. *Mon. Weather Rev.* **140**: 258–265, doi: 10.1175/MWR-D-10-05047.1.

Oke PR, Sakov P. 2008. Representation error of oceanic observations for data assimilation. *J. Atmos. Oceanic Technol.* **25**: 1004–1017.

Oke PR, Schiller A, Griffin DA, Brassington GB. 2005. Ensemble data assimilation for an eddy-resolving ocean model of the Australian region. *Q. J. R. Meteorol. Soc.* **131**: 3301–3311.

Ponte RM, Wunsch C, Stammer D. 2007. Spatial mapping of time-variable errors in Jason-1 and TOPEX/Poseidon sea surface height measurements. *J. Atmos. Oceanic Technol.* **24**: 1078–1085.

Rogel P, Weaver AT, Daget N, Ricci S, Machu E. 2005. Ensembles of global ocean analyses for seasonal prediction: Impact of temperature assimilation. *Tellus* **57A**: 375–386.

Schiller A, Oke PR, Brassington GB, Entel M, Fiedler R, Griffin DA, Mansbridge JV. 2008. Eddy-resolving ocean circulation in the Asian–Australian region inferred from an ocean reanalysis effort. *Prog. Oceanogr.* **76**: 334–365.

Silverman BW. 1986. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall: London.

Snyder C, Bengtsson T, Bickel P, Anderson J. 2008. Obstacles to high-dimensional particle filtering. *Mon. Weather Rev.* **136**: 4629–4640.

Thacker WC. 2003. Data-model-error compatibility. *Ocean Model.* **5**: 233–247.

van Leeuwen PJ. 2009. Particle filtering in the geosciences. *Mon. Weather Rev.* **137**: 4089–4114.

van Leeuwen PJ. 2010. Nonlinear data assimilation in geosciences: An extremely efficient particle filter. *Q. J. R. Meteorol. Soc.* **136**: 1991–1996.

van Leeuwen PJ. 2011. Efficient fully non-linear data assimilation in geophysical fluid dynamics. *Comput. Fluids* **46**: 52–58, doi: 10.1016/j.compfluid.2010.11.011.

Wu L, Bocquet M, Lauvaux T, Chevallier F, Rayner P, Davis K. 2011. Optimal representation of source-sink fluxes for mesoscale carbon dioxide inversion with synthetic data. *J. Geophys. Res.* **116**: D21304, doi: 10.1029/2011JD016198.

Zaron ED, Egbert GD. 2006. Estimating open-ocean barotropic tidal dissipation: The Hawaiian Ridge. *J. Phys. Oceanogr.* **36**: 1019–1035.